Les crawlers, des outils indispensables pour le SEO (2ème partie)

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

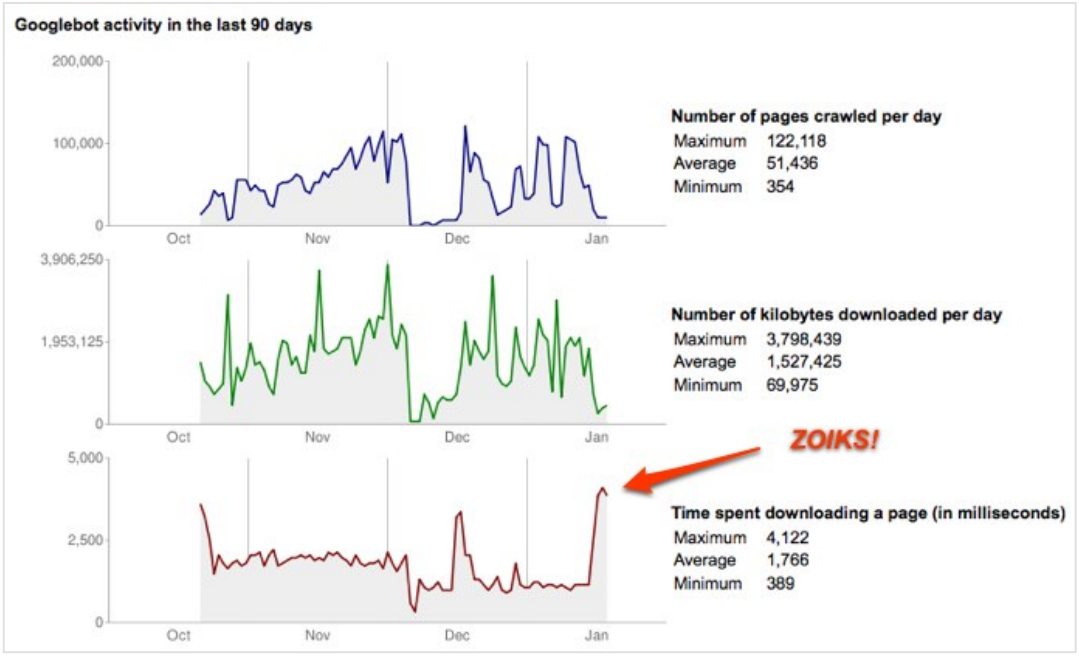
On connait tous les robots Googlebot et Bingbot, utilisés par leur moteur de recherche respectif pour explorer des milliards de sites web chaque jour. Mais il existe également des outils qui simulent ces robots et vous donnent de nombreuses indications sur l'indexabilité de votre site web. Des systèmes qui sont devenus indispensables au fil du temps dans le cadre d'un audit SEO. Après la découverte des différents outils disponibles ainsi que leur utilisation "classique" le mois dernier, nous nous attaquons à des usages plus avancés dans cette seconde partie...

Le mois dernier, nous avons passé en revue les utilisations « classiques » des crawlers. Mais ces outils peuvent être utilisés pour des analyses plus avancées, qui s'avèrent particulièrement utiles pour le SEO.

L'analyse des problèmes de performance

La fréquence et l'intensité des crawls de Google peuvent être fortement influencées par les performances de votre site web. Nous ne parlons pas ici du « temps de rendition », celui qui est expérimenté par l'utilisateur dans son navigateur, qui dépend de multiples facteurs, mais uniquement du délai nécessaire entre une requête "http://" et la réception de l'intégralité du code d'une page ou fichier.

On observe couramment qu'au-delà d'une seconde de temps de téléchargement, les bots de Google commence à ignorer les pages d'un site, et *a minima*, les crawlent moins souvent.

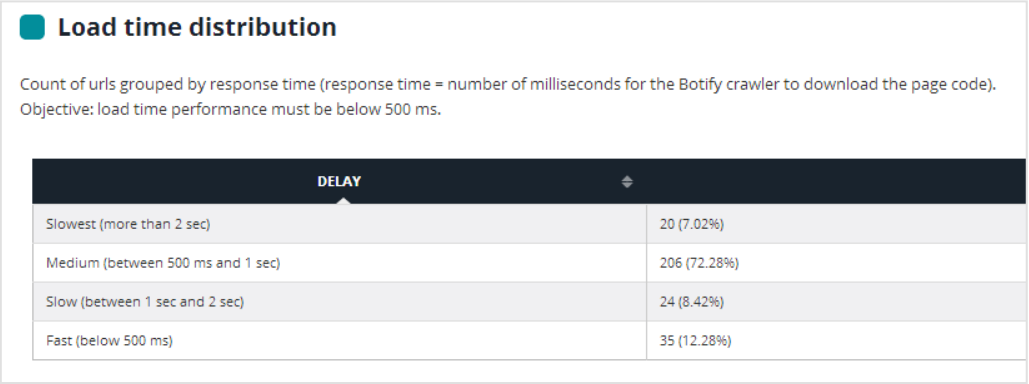


Un exemple typique des phénomènes constatés : ici le temps de téléchargement moyen indiqué dans les Webmaster Tools explose d'un seul coup, pour atteindre les 4 sec. Immédiatement, le nombre de pages crawlées tombe de 51000 pages par jour à 350.

Dans la pratique, trouver les causes de ces mauvaises performances se révèle souvent problématique. En effet, les outils de mesure de performance peuvent ne pas détecter ces anomalies, car ils se basent souvent sur des scénarios d'utilisation du site qui correspondent à des comportements d'utilisateurs et non de robots d'exploration. Or ces derniers ont une fâcheuse tendance à :
- Appeler des pages qui ne sont pas en cache (tout simplement parce qu'ils sont les premiers à appeler ces pages depuis le moment où la page en cache est devenue obsolète).
- A déclencher massivement des requêtes lourdes à calculer (comme celles correspondant à des pages de pagination).
- A appeler des pages dans un ordre qui n'est pas forcément celui « imaginé » par le développeur.
- Etc.

Google, *via* ses Webmaster Tools, ne fournit qu'un temps de téléchargement moyen : impossible de savoir quelles sont les pages qui sont lentes, et d'identifier le ou les scripts qui impactent les performances.

Une solution éprouvée consiste donc à utiliser un crawler qui mesure les temps de téléchargement, observés par un « bot » similaire à ceux de Google pour identifier les « motifs » (les « patterns ») d'URL plus lentes à télécharger que les autres. Plusieurs crawls, effectués à des heures différentes et avec des paramètres différents sont parfois nécessaires pour reproduire le problème mais en règle générale, cette méthode permet souvent d'isoler l'origine du défaut, et de trancher entre les différentes hypothèses (mauvais paramétrage du cache, requêtes non optimisées, code mal écrit…).

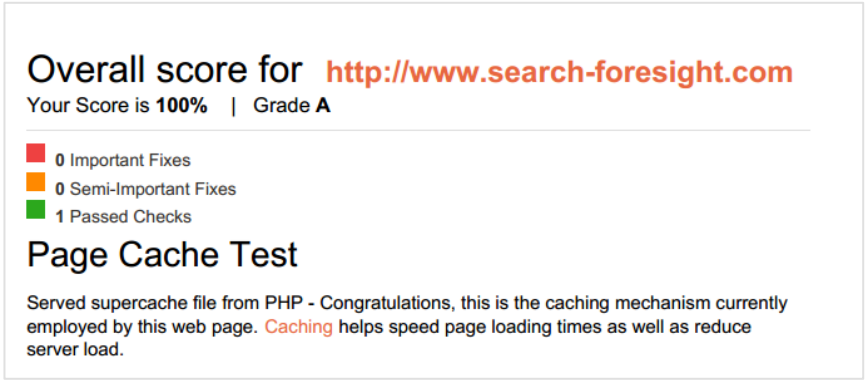


Load time distribution

Count of urls grouped by response time (response time = number of milliseconds for the Botify crawler to download the page code).
Objective: load time performance must be below 500 ms.

DELAY	
Slowest (more than 2 sec)	20 (7.02%)
Medium (between 500 ms and 1 sec)	206 (72.28%)
Slow (between 1 sec and 2 sec)	24 (8.42%)
Fast (below 500 ms)	35 (12.28%)

Exemple : temps de téléchargement mesurés par le crawler de Botify. Les crawlers mesurent souvent des temps supérieurs à ceux remontés par Google dans les webmaster tools, probablement parce que l'infrastructure de crawl de Google est très optimisée pour diminuer les temps de latence.



Overall score for http://www.search-foresight.com

Your Score is **100%** | Grade **A**

0 Important Fixes
0 Semi-Important Fixes
1 Passed Checks

Page Cache Test

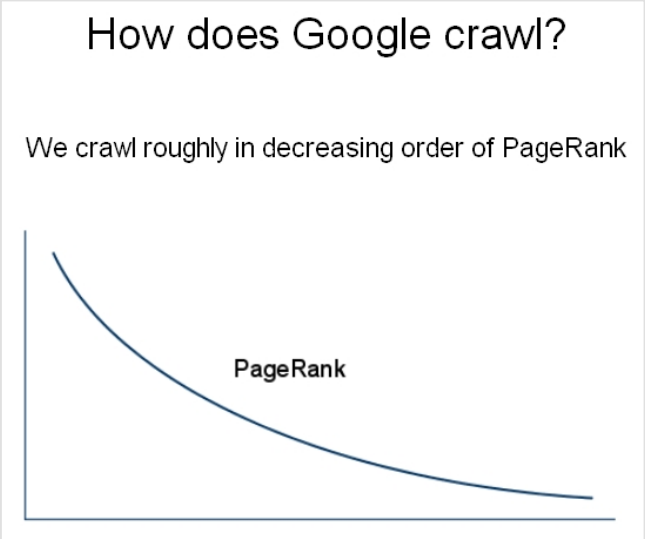
Served supercache file from PHP - Congratulations, this is the caching mechanism currently employed by this web page. Caching helps speed page loading times as well as reduce server load.

Test de fonctionnement du cache effectué à l'aide d'un crawler spécialisé. Souvent, les robots d'exploration créent involontairement des scénarios d'appel de pages pour lesquels le cache est inopérant ou le contenu caché est devenu obsolète.

Mesurer la profondeur des pages

Une autre utilisation particulièrement utile d'un crawler consiste à utiliser cet outil pour analyser la profondeur des URL. Ce que nous appelons profondeur dans ce contexte est le nombre de « sauts » (de clics) nécessaires pour atteindre une page en suivant les liens explorables, et en partant de la page d'accueil.

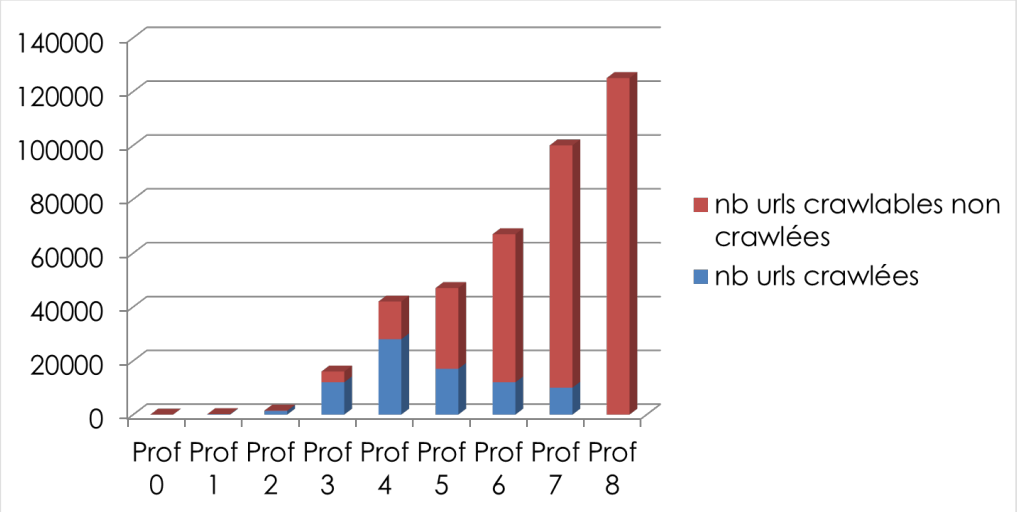
Le comportement de Google dépend de plusieurs facteurs. Parmi ces facteurs, l'un des plus importants est… le PageRank, comme Matt Cutts a déjà eu l'occasion de le préciser dans l'une de ses présentations au SMX en 2009 :



Le slide présenté par Matt Cutts sur le crawl :
plus le PR est élevé, plus Google aura envie de crawler cette page
et même de la recrawler souvent.

Or, compte tenu du facteur d'atténuation contenu dans la formule du Pagerank, plus une page est « profonde », moins elle a de chances de recevoir du « jus de lien » depuis la page d'accueil (qui est la page qui concentre toujours - sauf en cas d'arborescence bizarroïde - le plus de PageRank sur un site web). Le phénomène est accentué par le nombre croissant d'URL que l'on trouve sur un site au fur et à mesure que l'on descend dans son arborescence : plus la page est profonde, plus elle reçoit donc un PageRank dilué.

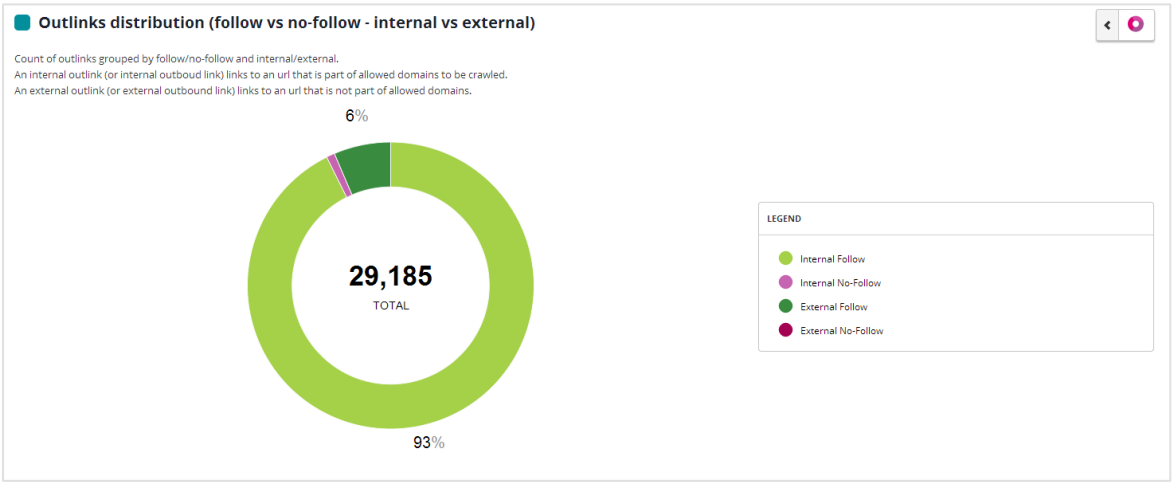
Mesurer la profondeur des pages à l'aide d'un crawler permet donc de vérifier que les pages « importantes » du site (celles qui ont un objectif SEO) ne sont pas « privées » de linkjuice à cause d'une arborescence mal conçue.



Analyse du nombre d'URL réellement crawlées par Google (information fournie par une analyse des logs serveurs en comptant les urls uniques appelées avec le user agent « Googlebot ») par rapport au nombre total d'URL du site en fonction de la profondeur. On voit ici qu'à partir de la profondeur 5, Googlebot commence à ignorer une majorité des URL, et aucune des très nombreuses URL situées en profondeur 8 n'est crawlée par le robot d'exploration de Google. Ce scénario est tout à fait commun sur des sites web comportant de nombreuses pages (tandis que les petits sites sont plus volontiers crawlés à 100%).

Mesurer l'indegree et l'outdegree des pages

Outre la profondeur des pages, la plupart des crawlers du marché sont également capables de fournir le nombre de liens sortants sur une page du site (l'*outdegree*), en séparant les liens pointant un domaine externe (*outdegree externe*) et les liens internes (*outdegree interne*).



Mesure de l'outdegree (appelé outlinks) sur l'outil Botify.
On peut aussi calculer les proportions de liens en dofollow ou en nofollow

La mesure de l'outdegree externe vs l'outdegree interne

Un ratio outdegree externe / outdegree interne révèle une fuite importante de PageRank depuis une page donnée (l'importance de cette fuite étant d'autant plus sérieuse que la profondeur de la page est faible par ailleurs).

La mesure de l'indegree vs l'outdegree

L'indegree est également un indicateur fourni par la plupart des crawlers du marché. Il mesure le nombre de liens entrants pointant vers une page.

Un ratio indegree/outdegree élevé révèle une page qui a tendance à accumuler du linkjuice, tandis que le contraire dénote une page qui transmet massivement du PageRank à d'autres pages.

La mesure de l'indegree

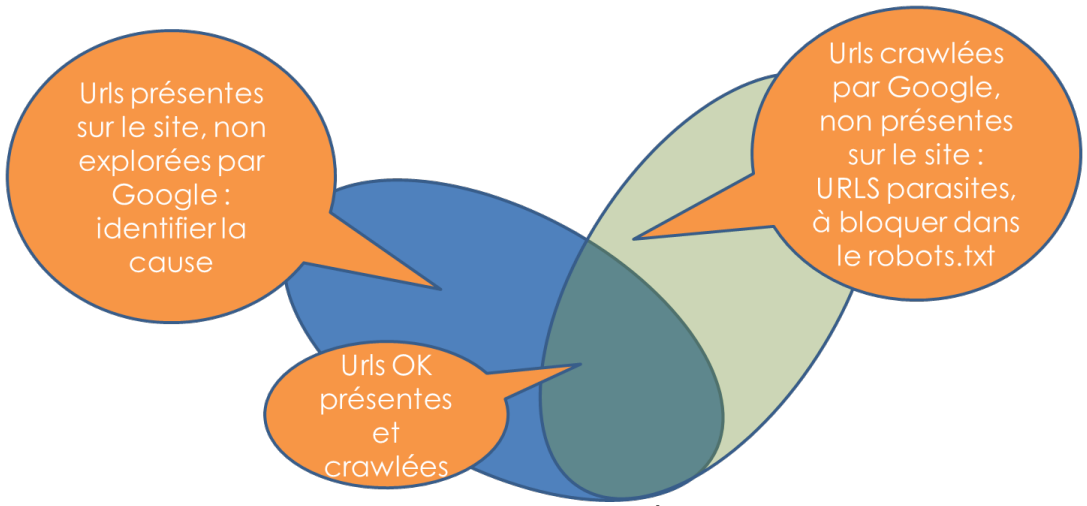
Les sites ont souvent une arborescence régulière, créée par l'utilisation de menus, de modèles de pages, et de système de catégorisation. Le nombre de liens sortants varie souvent assez peu d'une page à l'autre, et le nombre de liens entrants dépend avant tout de la structure des menus et de la hiérarchie des catégories. L'analyse du nombre de liens internes « entrants » sur les pages est par conséquent un bon indicateur pour savoir quelles sont les pages « favorisées » par une arborescence : ce sont celles qui ont l'indegree le plus élevé. Et une simple comparaison entre les pages à fort indegree et la liste des pages que vous souhaitez promouvoir révèle parfois des défauts importants dans la structure de maillage interne du site.

L'analyse combinée crawl + logs

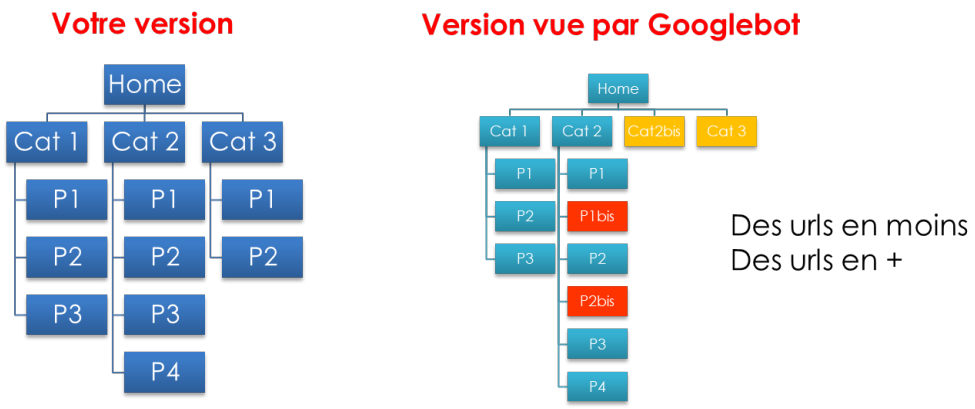
Si l'on dispose des logs serveurs de son site, et si l'on sait les analyser, on va pouvoir en tirer des informations qui, une fois croisées avec les données recueillies par le crawler, vont permettre un diagnostic avancé du « référencement » (indexation) du site.

En effet, en analysant les lignes des logs qui contiennent le user-agent « Googlebot », il est possible de savoir quelles sont les URL crawlées réellement par Googlebot (évidemment, on peut faire la même chose avec Bingbot et tous les autres « bots »).

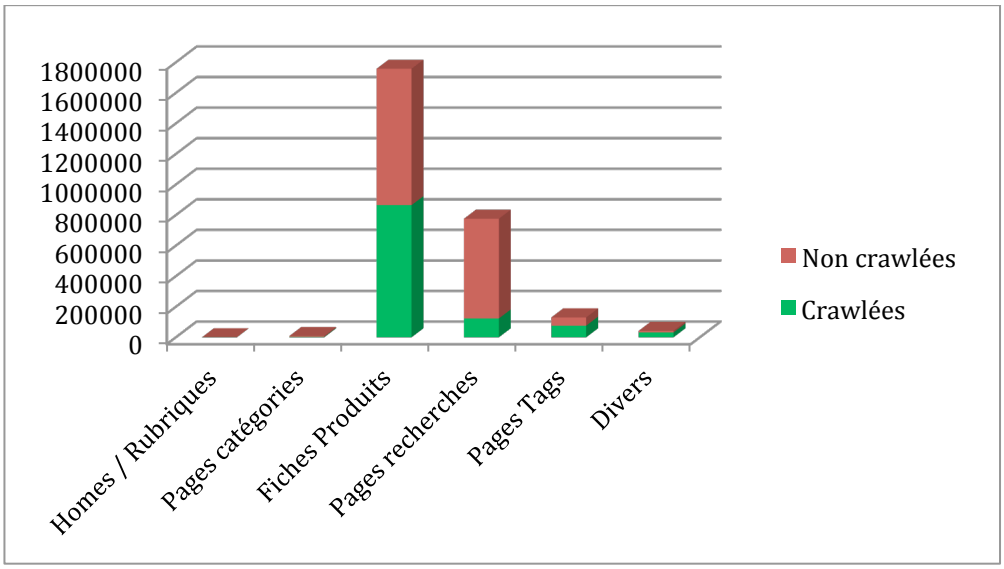
Une comparaison avec les données fournies par votre outil de crawl va donc permettre d'identifier :
- Les URL « explorables » (trouvées par votre crawler) mais que Googlebot n'explore pas.
- Les URLs « explorables » et réellement téléchargées par Googlebot.
- Les URL appelées par Googlebot et qui ne sont pas présentes sur les pages de votre site (et oui, il y en a toujours), et il n'est pas rare que les volumétries liées à ces cas soient très supérieures au nombre de pages figurant effectivement dans l'arborescence.



La comparaison des listes d'URL explorables (identifiées par votre crawler) avec les URL réellement crawlées par Google (identifiées dans les logs ou tout autre système enregistrant les « hits » de Googlebot) fournit immédiatement des chantiers SEO intéressants. Identifier la cause d'un « oubli » massif d'URL réellement présentes dans l'arborescence peut par exemple permettre des gains substantiels de trafic en provenance des moteurs de recherche.



L'arborescence vue par Googlebot peut être radicalement différente de la structure réelle du site. Si Googlebot appelle des variantes d'URL qui répondent un code 200 au lieu d'un code 404, il peut générer des « doublons » de type DUST (« different URL, same text ») qui n'existent pas sur le site, se perdre dans des pièges à robots (spider traps), ou générer massivement des pages à partir du moteur de recherche ou des facettes (filtres) dans un catalogue. Une simple erreur technique, même corrigée en l'espace de quelques heures, ou flux mal formé peut provoquer ce phénomène.



L'analyse des pages non crawlées par template permet de détecter que les problèmes de crawl affectent préférentiellement certains modèles de pages : dans cet exemple, les pages de recherches sont massivement ignorées par Google.

Les analyses avancées proposées par des crawlers spécialisés

Avec le temps, des fonctionnalités plus originales sont apparues sur certains crawlers, tandis qu'apparaissaient aussi des crawlers spécialisés.

Parmi ces derniers, on peut citer les crawlers dédiés à l'analyse et à la vérification des plans de taggage pour le web analytics. Mais aussi les crawlers dédiés à des fonctions de web mining ou de data mining (comme import.io par exemple).

Quant aux fonctionnalités originales, les plus intéressantes sont sans conteste les outils de détection de doublons ou de quasi doublons comme Siteliner qui fournissent des données difficiles à réunir en temps normal (la comparaison de pages oblige à recourir à des algorithmes avancés).

Le « crawler » est-il devenu le meilleur ami du référenceur ?

Dans la première partie, nous avons vu à quel point un crawler pouvait s'avérer particulièrement efficace pour détecter facilement des problèmes sur un site : liens brisés, URL anormales, codes d'erreurs inattendus renvoyés par l'applicatif, redirections étranges, régressions diverses en cas de nouvelles versions.

Pour un référenceur, il permet en outre de faire facilement des diagnostics sur la présence ou l'absence de balises SEO (title, H1, meta desc, meta robots, canonical, rel=next/prev, hreflang…) et sur les plaies que sont pour le référencement les problèmes de performance, les problèmes de profondeur et de maillage, mais aussi la présence de DUST ou de quasi doublons.

Dans la pratique, et en toute logique, leur usage chez les référenceurs s'est fortement démocratisé, et s'il existe parmi les lecteurs des webmasters et des référenceurs qui ne les utilisent pas encore, nous ne pouvons que leur conseiller de tester ceux décrits dans notre premier article : essayer les crawlers, c'est les adopter… Mais méfiez-vous : on en vient vite à crawler tout, tout le temps, et pour n'importe quel prétexte : c'est donc parfois « addictif » et par là-même coûteux lorsque les outils sont payants…

L'important est donc d'identifier quel outil utiliser, et pour quel objectif, et à la fin, vous vous rendrez compte que vos outils de crawls sont devenus vos meilleurs alliés pour contrôler et comprendre ce qui se passe sur vos sites.

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***